Expert Systems With Applications 80 (2017) 39–45

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Enhanced question understanding with dynamic memory networks for textual question answering

Chunyi Yue[a,b], Hanqiang Cao[a,*], Kun Xiong[c], Anqi Cui[c], Haocheng Qin[c], Ming Li[b]

[a] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, 430074, China
[b] David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, N2L 3G1, Canada
[c] RSVP Technologies Inc, Suite 19, 279 Weber St N, Waterloo, N2J 3H8, Canada

ARTICLE INFO

Article history:
Received 29 September 2016
Revised 1 March 2017
Accepted 2 March 2017
Available online 10 March 2017

Keywords:
Dynamic memory networks
Attention based GRU
Textual question answering

ABSTRACT

Memory networks show promising context understanding and reasoning capabilities in Textual Question Answering (Textual QA). We improve the previous dynamic memory networks to do Textual QA by processing inputs to simultaneously extract global and hierarchical salient features. We then use them to construct multiple feature sets at each reasoning step. Experiments were conducted on a public Textual Question Answering dataset (Facebook bAbI dataset) in two ways: with and without supervision from labels of supporting facts. Compared to previous works such as Dynamic Memory Networks, our models show better accuracy and stability.

1. Introduction

Automated reasoning is a field of artificial intelligence (AI). It connects with mathematical logic and computer science. Given some facts, the machine needs to conduct inferences and then make judgements from these facts. In Natural Language Processing (NLP), the task of Textual Question Answering (QA) can be seen as a type of reasoning tasks: Given a question, the machine provides an answer (judgement) based on a (miniature) knowledge base (facts) by analyzing the question, finding proper entities and attributes, and then retrieving the answer (inference steps). A sample of Textual QA is give in Fig. 1. The well-known intelligent system IBM Watson builds its knowledge base from many different sources, from encyclopedia to the Internet, from structured infoboxes to unstructured texts (Fan, Kalyanpur, Gondek, & Ferrucci, 2012). However, Textual QA has its difficulties: The facts are finite, simple sentences involving several objects (entities). To answer a question, the machine must infer from the single source of limited facts precisely and recognize the entities and relations accurately. Though challenging, AI researchers have built inference engines as components in expert systems (Jackson, 1998), to deduct new knowledge from existing knowledge bases. Typically the inference engines work with logics represented as IF-THEN rules, constructed from explicit variables, predicates and quantifiers. However, for natural language understanding, parsing the sentence may be difficult; the noises introduced may collapse the fragile logical system.

The recent success of deep neural networks has brought a new solution to this traditional task. Firstly proved by some image processing tasks, the neural networks have shown great potential of capturing connections between the observed elements, i.e., pixels or words (Antol et al., 2015; Vinyals, Toshev, Bengio, & Erhan, 2015; Xu, Courville, Zemel, & Bengio, 2015; Yang, He, Gao, Deng, & Smola, 2015). In NLP, the structures of convolutional neural networks (CNN) (Hu, Lu, Li, & Chen, 2014) and recurrent neural networks (RNN) (Sutskever, Vinyals, & Le, 2014) map the words to higher dimensions while keeping tracks of their contexts, hence are effective in many classification tasks (Kim, 2014; Lai, Xu, Liu, & Zhao, 2015) and sequential tasks (e.g.machine translation) (Cho, van Merrienboer, Bahdanau, & Bengio, 2014; Dzmitry Bahdanau & Kyunghyun, 2015; Kalchbrenner & Blunsom, 2013; Meng, Lu, Tu, Li, & Liu, 2016). In addition, Memory Networks (Weston, Bordes et al., 2015; Weston, Chopra, & Bordes, 2015) and Neural Turing Machines (Graves, Wayne, & Danihelka, 2014) introduce external memory units and flexible information storage mechanisms.

The paper is organized as follows. After reviewing related work in Section 2, we present our basic model in Section 3. Compared to other dynamic memory networks (DMNs) (Kumar et al., 2016; Xiong, Merity, & Socher, 2016), our basic model has a subtly different internal structure and Attention based GRU (AttenGRU) (Kumar et al., 2016; Xiong et al., 2016) mechanism. In Section 4, we present our improved model - EnDMN. In Section 5, we show and analyze experimental results. In Section 6, we summarize the main contributions of this work and propose some future research.

[*] Corresponding author.
E-mail addresses: d201377468@hust.edu.cn (C. Yue), caohq@hust.edu.cn (H. Cao), kun@rsvptech.ca (K. Xiong), caq@rsvptech.ca (A. Cui), qhc@rsvptech.ca (H. Qin), mli@uwaterloo.ca (M. Li).

http://dx.doi.org/10.1016/j.eswa.2017.03.006
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

Facts	*supporting*	+	Question	→	Answer
Lily is a swan.					
Bernhard is a lion.	*Yes* 2				
Greg is a swan.					
Bernhard is white.	*Yes* 3				
Brian is a lion.	*Yes* 1	+	What color is Brian?	→	White
Lily is gray.					
Julius is a rhino.					
Julius is gray.					
Greg is gray.					

Fig. 1. A sample from Facebook bAbI dataset.

2. Related work

Open-domain Question Answering (openQA) is a classical QA task, which requires an intelligent system to directly output a precise answer in natural language after receiving a question. For example, when a user inputs the question *"what is the largest inland lake in China?"*, the system is expected to output an answer *"Qinghai Lake"* rather than a list of ranked snippets and links. Recently, an increasing number of knowledge bases (e.g., Freebase, YAGO, and Google Knowledge) and corpus bases (e.g., blogs and forums) have become accessible. Combining with other techniques, some new progress in openQA has been obtained (Bhati & Prasad, 2016; Sun et al., 2015).

Community Question Answering (CQA) is another hot issue in the field of QA. Many online CQA platforms (e.g., Quora and Stack Overflow) have become popular, where users can share knowledge in an interactive way. There is an obvious advantage to CQA - it allows users to obtain expected knowledge from other users in a variety of ways. Users can post their questions or answers, and comment or vote on contents posted by other users in the community. Namely, a user can be a questioner, an answerer, or a reviewer. Recently, many improved CQA systems have been proposed (Chang & Pal, 2013; Sahu, Nagwani, & Verma, 2016a; 2016b).

Question classification (QC) is a key part of traditional search engines and QA systems. QC can determine the entity of an answer and the pattern of an answer beforehand, which reduces the search scope and promotes search accuracy for the following information retrieval and answer selection. For instance, given questions *"who was the first man to win the Nobel Prize in Literature?"* and *"what is a violin?"*, the answer to the first question is supposed to be a name, and the answer to the second question should have a particular pattern like *"A violin is..."* or *"The violin is..."*. Some hybrid approaches to improve the performance of QC can be found in Loni (2011).

As described above, there is a significant difference between Textual QA and other QA tasks. In Textual QA, the question is always closely related to a miniature knowledge base (facts) about a particular scene. In this work, we focus on Textual QA. The main contributions of our work are threefold. First, we introduce global and hierarchical salient features of inputs (a question and a series of facts). Other models only use one type of features. Second, we propose using a modified network to extract the hierarchical salient features of a question to further improve the overall performance of our model. Third, we find a method to utilize these features to control the extraction of the information at each reasoning step. A main difference between our model and the closest related approaches is shown in Table 1.

Table 1
A main difference between our model and the closest related models.

EnDMN	DMNs
Global feature set **and** salient feature set	Global feature set **or** salient feature set

3. Basic framework and approach

As mentioned in Section 1, logical rules and the chaining mechanism are the traditional methods in building expert systems and solving logical problems, as well as the succeeding work of semantic networks with ontology. These strategies involve manual organization and labeling which are costly and time-consuming, hence are unsuitable to make use of the big amount of data.

The advance of deep learning revolution has presented new hope. With the development of memory networks and their attention mechanisms, some logical reasoning tasks have become popular and practicable recently. Researchers do not have to build the knowledge base (ontology) themselves; but instead can solve Textual QA tasks with end-to-end neural networks such as End-To-End Memory Network (E2E) (Sukhbaatar & Szlam, 2015), Dynamic Memory Networks (DMN) (Kumar et al., 2016), Dynamic Memory Networks for Visual and Textual Question Answering (DMN+) (Xiong et al., 2016), Neural Reasoner (NR) (Peng, Lu, Li, & Wong, 2015) and so on (Andreas, Rohrbach, Darrell, & Klein, 2016; Yu, Zhang, Hang, & Zhou, 2015).

3.1. DMN

A DMN is a type of end-to-end neural networks. It is usually composed of four modules: an input module, a question module, an episodic module, and an answer module. There are various networks to choose from for each module in a DMN. In this paper, our basic model is assembled as below:

Input module: The input module is seen as sentence readers to process facts. Different encoding methods, including long-short term memory (LSTM) (Hochreiter & Schmidhuber, 1997), gated recurrent units (GRUs) (Kumar et al., 2016; Xiong et al., 2016), and position encoding (PE) (Sukhbaatar & Szlam, 2015), are usually applied in a sentence reader. It contains two parts. The first part is a PE (Sukhbaatar & Szlam, 2015) layer, which is used to produce original representations of facts s_i by:

$$s_i = \sum_j (l_j \cdot Ax_{ij}) \tag{1}$$

Where A is a word embedding matrix, l_j is a column vector composed of the elements $l_{kj} = (1 - j/J) - (k/d)(1 - 2j/J)$, with the number of words in the sentence J and the dimension of the word embedding d. The second part is a bidirectional gated recurrent neural network (Chung, Gulcehre, Cho, & Bengio, 2014; Schuster & Paliwal, 1997), which is used to produce final representations of facts $\overleftrightarrow{f_i}$. A same structure with different parameters is adopted to produce another final representations of facts $\overleftrightarrow{f_i^{(a)}}$, which are used to produce attention weights in the episodic memory module.

Question module: The question module of our basic model also is a sentence reader to process a question. It includes a Recurrent Neural Network (RNN). The final hidden state is seen as the representation of a question q.

Episodic memory module: This module is a core part of a DMN, where the input module interacts with the question module. Typically, a DMN uses a recurrent attention structure to achieve the progressive information extraction or reasoning in the episodic memory module. A reasoning step is regarded as a hop. There are two mechanisms in the episodic memory module: an attention mechanism (Luong, Pham, & Manning, 2015) and a memory